ترجمه ماشینی مبتنی بر مثال با کمک هستان‌شناسی

حسن شجاعی مند
دانشجوی کارشناسی ارشد نرم‌افزار

خلاصه

در سال‌های اخیر، یک روش ترجمه ماشینی جدیدی با عنوان ترجمه ماشینی مبتنی بر مثال، در مقالات مختلف مطرح شده است. در این روش، سیستم حاوی یک پایگاه دانش می‌باشد. با ورود یک جمله جدید برای ترجمه، سیستم از همین عبارات موجود که با یکدیگر منطبق شده‌اند، کمک می‌گیرد تا ترجمه عبارت جدید را بدست آورد. تجزیه جملات به عبارات جزئی‌تر و چیدمان آنها، می‌تواند کارایی سیستم را افزایش دهند. در این سیستم، می‌توان از موارد مطرح در وب معنایی، برای ذخیره‌سازی اطلاعات موردنیاز در پایگاه دانش، برای چیدمان‌های مورد نیاز در مرحله تشکیل پایگاه دانش و چیدمان‌های مورد نیاز در مرحله انجام تطابق، استفاده نمود. در این مقاله، روشی مبتنی بر استفاده از هستان‌شناسی، برای یک سیستم ترجمه ماشینی مبتنی بر مثال ارائه می‌شود.

کلمات کلیدی: ترجمه ماشینی مبتنی بر مثال، وب معنایی، هستان‌شناسی

۱- مقدمه

واژه ترجمه ماشینی، به سیستم‌های کامپیوتری اطلاق می‌شود که متنی را از یک زبان به زبان دیگر به صورت اتوماتیک ترجمه می‌کنند. هرچندکه هدف اصلی یک سیستم ترجمه ماشینی، تولید یک ترجمه با کیفیت بالا بدون دخالت انسان می‌باشد، در عمل اینکار تنها در موقعیت‌های بسیار محدود شده و با کیفیتی بسیار پایین امکان‌پذیر شده است. سیستم‌های ترجمه ماشینی، روش‌های مختلف موجود و موارد کلی مطرح در این سیستم‌ها، در کتاب‌های [1,2] مورد بررسی قرار گرفته است. به صورت کلی، سیستم‌های ترجمه ماشینی را می‌توان به دو دسته کلی، سیستم‌های مبتنی بر قاعده و سیستم‌های مبتنی بر یک پایگاه‌دانش تقسیم‌بندی کرد. در سیستم‌های مبتنی بر قاعده، سعی بر آن است تا با استفاده از قواعد زبانی موجود در زبان مبدا و مقصد، و استفاده از روابط موجود بین این قواعد زبانی، ترجمه یک جمله ورودی را به زبان مقصد بدست آوریم. در این روش، عدم امکان تعیین دقیق قواعد زبانی و روابط بین قواعد زبانی مختلف و همچنین ابهام‌هایی که در ترجمه کلمات از یک زبان به زبان دیگر وجود دارد و معانی مختلفی که کلمات می‌توانند در ترجمه داشته باشند، از مشکلات اصلی محسوب می‌شوند. در سال‌های اخیر، روش‌های مبتنی بر پایگاه دانش، معرفی شده‌اند. این روش‌ها، به دو دسته کلی سیستم‌های ترجمه ماشینی مبتنی بر مثال و سیستم‌های ترجمه ماشینی مبتنی بر آمار تقسیم‌بندی می‌شوند. ایده اصلی در آنها، استفاده از ترجمه‌های قبلی انجام شده برای رسیدن به ترجمه یک جمله جدید می‌باشد. در بخش۲، یک مرور کلی خواهیم داشت بر سیستم ترجمه ماشینی مبتنی بر مثال خواهیم داشت و مواردی که در این سیستم‌ها مطرح می‌باشد. در بخش۳، کارهای انجام شده در خصوص این سیستم‌ها و مرتبط با وب معنایی را مرور خواهیم کرد، در بخش۴، سیستم پیشنهادی این مقاله را معرفی خواهیم کرد و در بخش ۵، سیستم را با چند مثال ارزیابی خواهیم کرد.

۲- ترجمه ماشینی مبتنی بر مثال[1]

ایده ترجمه ماشینی مبتنی بر مثال، به سال ۱۹۸۴ بر می‌گردد که در مقاله [10] به صورت زیر معرفی شده است:

افراد ترجمه یک جمله را با استفاده از انجام تحلیل‌های زبانی عمیق انجام نمی‌دهند، بلکه ابتدا آنرا به یکسری عبارات می‌شکنند و سپس این عبارات را به زبان مقصد ترجمه کرده و سپس عبارات بدست‌آمده را با یکدیگر ترکیب می‌کنند تا جمله نهایی در زبان مقصد بدست آید. ترجمه عبارات بدست آمده نیز با استفاده از اصل ترجمه قیاسی[2] با مثال‌های مشابه ترجمه‌شده قبلی، انجام خواهد شد.

در همان مقاله [10]، سه قسمت اصلی برای یک سیستم ترجمه ماشینی مبتنی بر مثال، در نظر گرفته شده است: مطابقت دادن عبارات جزئی بدست‌آمده با یک بانک‌اطلاعاتی حاوی مثال‌های ترجمه شده، تشخیص ترجمه عبارات جزئی، اتصال دوباره عبارت جزئی برای بدست‌آوردن جمله نهایی.

برای روشن‌شدن این روش ترجمه، از یک مثال استفاده می‌کنیم. ترجمه جمله (1) را می‌توان با استفاده از ترجمه عبارات مشخص شده در جملات (2) و (3) که زیر آنها خط کشیده شده است، به صورت (4) بدست‌آورد.

(۱) He buys a book on semantic web.

(۲) He buys a notebook.

او یک دفترچه یادداشت می‌خرد.

(۳) I read a book on semantic web.

من یک کتاب درباره وب‌معنایی می‌خوانم.

(۴) او یک کتاب درباره وب‌معنایی می‌خرد.

با توجه به مثال عنوان شده فوق و با استفاده از [8,3,13,14]، که یک مرور کامل بر روی سیستمهای ترجمه ماشینی مبتنی بر مثال انجام داده است، اجزاء اصلی این سیستم عبارتند از: مجموعه مثال‌ها[3]، تطابق[4] و ترکیب‌مجدد[5]. هر کدام از این اجزاء اصلی و کارهای انجام شده در خصوص آنها را در ادامه بیشتر بررسی می‌کنیم.

- مجموعه مثال‌ها

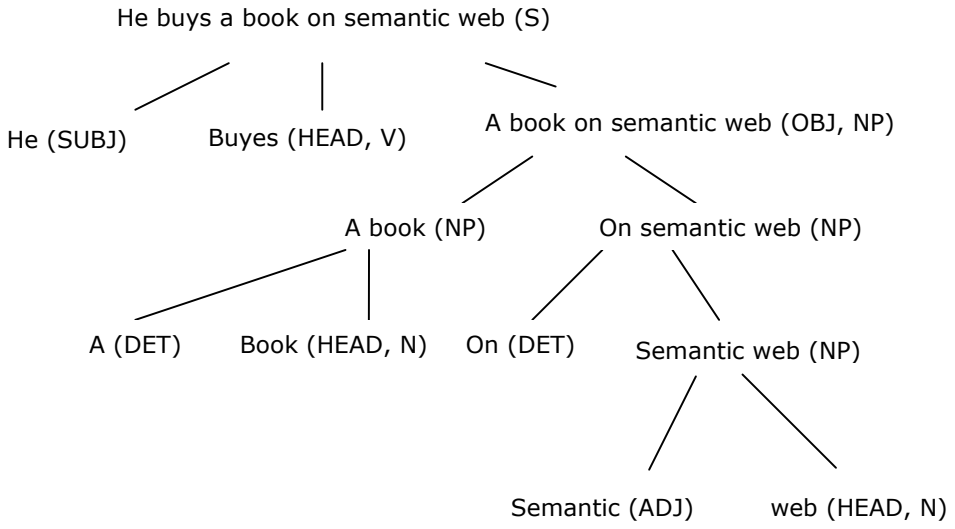
این سیستم، نیاز به یک بانک‌اطلاعاتی دارد که حاوی جملات به زبان مبدا و ترجمه آنها به زبان مقصد باشد. برای جملات ترجمه شده معمولا از ترجمه‌هایی که به صورت دستی و توسط یک عامل انسانی انجام شده است، استفاده می‌کنند. سیستم‌های ترجمه‌ماشینی مبتنی بر مثال، از روش‌های مختلفی برای ذخیره مثال‌ها استفاده می‌کنند. واضح است که، روشی که یک سیستم برای ذخیره مثال‌ها استفاده می‌کند، تاثیر مستقیمی بر روی جستجوی مثال‌های مشابه برای یک جمله ورودی دارد. در ساده‌ترین حالت، مثال‌ها به صورت جفت رشته‌هایی بدون هیچ‌گونه اطلاعات اضافی ذخیره می‌شوند. و در حالتی که تعداد مثال‌ها زیاد باشد، می‌توان از یکی از روش‌های ایندکس‌گذاری استفاده کرد.

یکی از معمول‌ترین روش‌های ذخیره مثال‌ها، استفاده از ساختار درختی می‌باشد. برای تشکیل درخت، از روش تحلیل وابستگی[6] استفاده می‌شود که در [12] ارائه شده است. تحلیل وابستگی، با استفاده از ساختار گرامری جمله، درختی را برای آن ارائه می‌کند. این فرایند ، جمله را به اجزای کوچکتری بر اساس قواعد گرامری موجود می‌شکند و این‌کار را تا

---

[1] Example based machine translation
[2] Analogy translation principle
[3] Parallel corpora
[4] matching
[5] recombination
[6] Dependency analysis

آنجا ادامه می‌دهد که به اجزاء ریز گرامری برسیم. درخت وابستگی ساخته‌شده، به ازای هر گره، یکسری اطلاعات در خصوص کلمه، نوع‌گرامری و ... را دارا می‌باشد. یک نمونه درخت وابستگی ایجاد شده، در شکل۲ نشان داده شده است.



شکل۱: درخت وابستگی

یکی از مواردی که در کارایی این سیستم ترجمه ماشینی تاثیر می‌گذارد، ریزدانگی[1] مثال‌های موجود در بانک‌اطلاعاتی می‌باشد. هرچقدر طول مثال‌ها بزرگتر باشد، احتمال یافتن یک انطباق کامل در مجموعه مثال‌ها، کاهش می‌یابد و هر چقدر که طول مثال‌ها کوچکتر باشد، احتمال ابهام افزایش خواهد یافت، زیرا برای یک عبارت چندین انطابق در مجموعه مثال‌ها پیدا خواهیم کرد و همچنین کیفیت ترجمه کاهش خواهد یافت. تجزیه جمله به عبارات کوچکتر، با مشکل چیدمان اجزای کوچکتر مواجه است، که ممکن است این چیدمان انجام شده دقت کافی را نداشته و در نتیجه بر کیفیت ترجمه نهایی تاثیر بگذارد. در [9] این مسئله برای سیستم‌های ترجمه ماشینی مبتنی بر آمار که در بخش پایگاه دانش با سیستم‌های ترجمه ماشینی مبتنی بر مثال یکسان هستند، مورد بررسی قرار گرفته است.

تعداد مثال‌های موجود در بانک‌اطلاعاتی نیز، از دیگر موارد تاثیرگذار بر کارایی سیستم ترجمه ماشینی می‌باشد. تعداد مثال‌ها، بستگی به زبان‌ها و حوزه‌ای دارد که می‌خواهیم عمل ترجمه را انجام دهیم. به طور معمول، هرچقدر تعداد مثال‌ها افزایش یابد، کیفیت ترجمه نیز افزایش خواهد یافت. ولی بعد از اینکه تعداد مثال‌ها از یک حدی افزایش یابد، کارایی سیستم کاهش خواهد یافت و بنابراین باید تعادلی بین کیفیت موردنیاز و کارایی سیستم با انتخاب مثال‌های مناسب برای سیستم صورت گیرد. در [3] لیستی از سیستم‌های ترجمه‌ماشینی مبتنی بر مثال، به همراه تعداد مثال‌هایی که در بانک‌اطلاعاتی آنها وجود دارد، ارائه شده است.

در [7]، با این فرضیه که بسیاری از اجزای ترجمه‌ماشینی دوباره در متن‌های دیگر وجود داشته و نیاز به ترجمه دارند، روش‌هایی برای تشخیص این اجزای تکراری و اولیت‌بندی آنها برای ترجمه‌های آتی، ارائه شده است.

- تطابق

پس از دریافت جمله ورودی، سیستم باید از بانک‌اطلاعاتی مثال‌هایی را که به جمله ورودی نزدیک هستند، را استخراج نماید. این مرحله از یک سیستم ترجمه ماشینی، به نحوه ذخیره‌سازی مثال‌ها در بانک‌اطلاعاتی وابسته است. در تمامی

[1] granularity

روش‌های تطابق، از مقیاس‌های اندازه‌گیری میزان شباهت یا تمایز استفاده می‌شود. در ساده‌ترین حالت که مثال‌ها به صورت رشته‌هایی ذخیره می‌شوند، می‌توان از روش‌های معمول محاسبه میزان شباهت یا تمایز دو رشته استفاده نمود، که شامل تطابق مبتنی بر کاراکتر، تطابق مبتنی بر کلمه، استفاده از روش زاویه شباهت[1] و دیگر روش‌های اندازه‌گیری میزان شباهت یا تمایز دو رشته استفاده نمود. در مواردی که علاوه بر رشته‌ها، یکسری اطلاعات اضافی در خصوص جملات نگهداری می‌شود، می‌توان از روش‌های دیگری مانند روش تطابق مبتنی بر کلمه حاشیه‌نویسی شده، نیز استفاده کرد. در این روش، علاوه بر در نظر گرفتن کلمات موجود، ساختار نحوی آنها در جمله نیز مورد توجه قرار می‌گیرند.

- ترکیب مجدد

بعد از یافتن مثال‌های مشابه، در مرحله بعد باید با استفاده از این مثال‌ها و ترجمه‌آنها، عبارات جزئی مورد نیاز را از مثال‌های مشابه استخراج کرده و برای بدست آوردن ترجمه نهایی آنها را با یکدیگر ادغام کنیم. در صورتیکه، مثال‌ها را با استفاده از یک ساختار درختی که عبارات جزئی آنرا از یکدیگر مجزا کرده و برچسب‌گذاری شده است، استفاده کرده باشیم، تشخیص عبارت جزئی مثال‌ها که تشکیل‌دهنده جمله ورودی می‌باشند، چندان مشکل نخواهد بود.

- ارزیابی سیستم

با در نظر گرفتن جدول شماره ۱، دو فاکتور کلی وجود دارد که در نحوه ارزیابی سیستم‌های ترجمه‌ماشینی دخالت دارند:

۱- مسلما مطمئن‌ترین ترجمه و باکیفیت‌ترین آن، ترجمه‌ای است که توسط انسان انجام می‌شود. از آنجاییکه ترجمه‌های متفاوتی از یک متن می‌تواند توسط انسان‌های مختلف انجام شود، بنابراین تعیین یک استاندارد طلایی[2] برای مجموعه تست مورد نیاز برای ارزیابی، به‌روشنی قابل انجام نیست.

۲- از آنجاییکه در ترجمه امکان جابجایی زیادی در محل کلمات بکاررفته در جمله داریم، تعیین فاصله بین جمله ترجمه‌شده توسط ماشین و جمله ترجمه‌شده توسط انسان، به‌روشنی قابل اندازه‌گیری نیست.

معدل علی در ترم قبل ۱۴ شد.
علی در ترم قبل، معدلش ۱۴ شد.
در ترم قبل، معدل علی ۱۴ شد.
علی در ترم قبل، معدل ۱۴ بدست آورد.

جدول۱: ترجمه‌های انسانی از یک جمله انگلیسی

تاکنون، محققان رشته ترجمه‌ماشینی نتوانسته‌اند ایده‌های خود را در زمینه ارزیابی سیستم ترجمه‌ماشینی معتبرسنجی نمایند، که این بدلیل همان طبیعت مفهومی بودن فرایند ارزیابی می‌باشد. روش‌های ارزیابی پیشنهاد شده، حاوی روش‌هایی شامل شمارش میزان خطاهای لغوی، نحوی و معنایی می‌باشد[3]. در [15] یک روش اتوماتیک برای ارزیابی سیستم‌های ترجمه ماشینی ارائه شده است. روش‌های انسانی برای ارزیابی سیستم‌های ترجمه ماشینی، کند و گران است و از طرفی مدت زمان زیادی لازم دارد. روش ارائه شده با نام BLEU، از معیار شباهت n-gram استفاده می‌کند که از یکسری جملات کاندید به همراه احتمال‌های مختلف ترجمه، استفاده می‌کند. این روش بر روی ۵ سیستم تست شده است و نتایج آنگونه که در همان مقاله آمده است، بسیار به ارزیابی‌های انسانی نزدیک بوده است.

[1] Angle of similarity
[2] Gold standard

۳- کارهای مرتبط

ایده وب‌معنایی، در [4] مطرح شد. وب‌معنایی، توسعه‌ای بر وب فعلی محسوب می‌شود که در آن معانی نیز در نظر گرفته شده‌اند، به‌گونه‌ای که کامپیوترها و انسان‌ها بتوانند همکاری بهتری با یکدیگر داشته باشند. در [5]، استفاده از روش‌های حاشیه‌نویسی که در وب‌معنایی مطرح می‌شود و کاربرد آنها در سیستم ترجمه ماشینی مبتنی بر مثال، مورد بررسی قرار گرفته است. یکی از موارد کاربرد روش‌های حاشیه‌نویسی در وب مانند RDF، استفاده از این صفحات وبی است که به دو زبان در وب موجود می‌باشند. RDF این امکان را فراهم می‌کند تا بتوان جملات مختلف و ترجمه‌های متناظر آنها را از آن صفحات استخراج کرده و از آنها در سیستم ترجمه ماشینی مبتنی بر مثال، استفاده نمود. همچنین می‌توان از توانایی RDF در ثبت معانی مورد نیاز برای مثال‌های موجود در بانک‌اطلاعاتی سیستم، استفاده نمود. از این طریق، می‌توان ابهام‌های موجود در ترجمه‌ها را که نیاز به معانی جمله دارند، برطرف نمود. یک روش ابتدایی برای استفاده از RDF در روش ترجمه ماشینی مبتنی بر مثال، در مقاله [6] ارائه شده است. استفاده از هستان‌شناسی در آن، در ابتدای فرایند سیستم و قبل از یافتن انطباق‌های جمله از بانک‌اطلاعاتی انجام می‌شود. معانی مورد نیاز جمله ورودی، از هستان‌شناسی استخراج شده و در یافتن انطباق‌های دقیق‌تر مورد استفاده قرار می‌گیرد.

۴- سیستم پیشنهادی

۵- ارزیابی سیستم

۶- جمع‌بندی و نتیجه‌گیری

مراجع

[1] D.J. Arnold, Lorna Balkan, Siety Meijer, R.Lee Humphreys and Louisa Sadler. *Machine Translation: an Introductory Guide*, Blackwells-NCC, London, 1994

[2] Daniel Jurafsky & James H. Martin. Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition. Draft of August 3, 2006.

[3] H. Somers. Review Article: Example-based Machine Translation. 2001.

[4] Tim Berners-Lee. Semantic Web Road map, 1998. http://www.w3.org/DesignIssues/Semantic.html

[5] Cristina Vertan, "Language Resources for the Semantic Web - Perspectives for Machine Translation", Proceedings of the Second International Workshop on Language Resources for Translation, work, research and Training, Coling '04 pg. 37 – 42. 2004.

[6] Natalia Elita & Antonina Birladeanu: A first step in integrating an EBMT into the Semantic Web MT Summit X, Proceedings of Workshop on Semantic Web Technologies for Machine Translation; pp.13-15. Phuket, Thailand, September 12, 2005.

[7] G. Craig Murray, Bonnie J. Dorr, Jimmy Lin, Jan Hajič, & Pavel Pecina: Leveraging recurrent phrase structure in large-scale ontology translation. EAMT-2006: 11th Annual Conference of the European Association for Machine Translation, Oslo, Norway. Proceedings; p.141-150, June 19-20, 2006.

[8] K. Knight, D. Marcu. Machine Translation in the Year 2004, Proc. ICASSP, 2005.

[9] Y. Deng, S. Kumar, and W. Byrne. Bitext Chunk Alignment for Statistical Machine Translation. CSLP Tech Report, Johns Hopkins University, 2004.

[10] Nagao, M.: 'A Framework of a Mechanical Translation between Japanese and English by Analogy Principle', in A. Elithorn and R. Banerji (eds), Artificial and Human Intelligence, Amsterdam: North-Holland, pp. 173–180. 1984.

[11] Consortium KnowledgeWeb : State of the Art on Ontology Alignment. Deliverable 2.2.3, FP6-507482, 2004.

[12] Broker, Norbert and Greet-Jan Kruijff. Dependency Grammer. 1999.

[13] Bowker, L. Computer-Aided Translation Technology. A Practical Introduction. Ottawa: University of Ottawa Press. 2002.

[14] A-Way, and M. Carl. "Introduction to Example-based machine Translation", Kluwer Academic Press, 2003.

[15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. 40th Annual Meeting of the ACL, pages 311--318, Philadelphia, PA, Jul. 2002.